Introduction

Big Data Analytics

Presented by: Dr Sherin El Gokhy

Adv. Methods

Module 4 – Advanced Analytics - Theory and Methods

Module 4: Advanced Analytics – Theory and Methods

Part 3: Linear Regression

During this Part the following topics are covered:

- General description of regression models
- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The Reasons to Choose (+) and Cautions (-) of the linear regression model

Regression

- Regression focuses on the relationship between an outcome and its input variables.
- Regression analysis helps one understand how the value of the dependent variable (also referred to as outcome) changes when any one of the independent (or input) variables changes, while the other independent variables are held fixed.
  - Provides an estimate of the outcome based on the input values (simply predict the outcome).
  - Models how changes in the input variables affect the outcome.
- The outcome can be continuous or discrete.
- Data scientists apply regression techniques as ***predictors or classifiers***.
- Possible use cases:
  - Estimate the lifetime value (LTV) of a customer and understand what influences LTV…..Linear regression
  - Estimate the probability that a loan will default or not and understand what leads to default….Logistic regression
- **Our approaches: linear regression and logistic regression**

Linear Regression

- Used to estimate a continuous value as a linear and additive function of other variables
  - Income as a function of years of education, age, and gender
  - House sales price as function of square footage, number of bedrooms/bathrooms, and lot size
- Outcome variable is continuous.
- Linear regression is **a commonly used technique for modeling a continuous outcome**.
- Input variables can be continuous or discrete.
- Model Output:
  - A set of estimated coefficients *that indicate the relative impact of each input variable on the outcome*
  - A linear expression for estimating the outcome as a function of input variables

Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon$$

- In linear regression, the outcome variable is expressed as a linear combination of the input variables (a weighted sum of input variables).
- where y is the outcome variable

  x_j are the input variables, for *j* = *1,2,…,p−1*

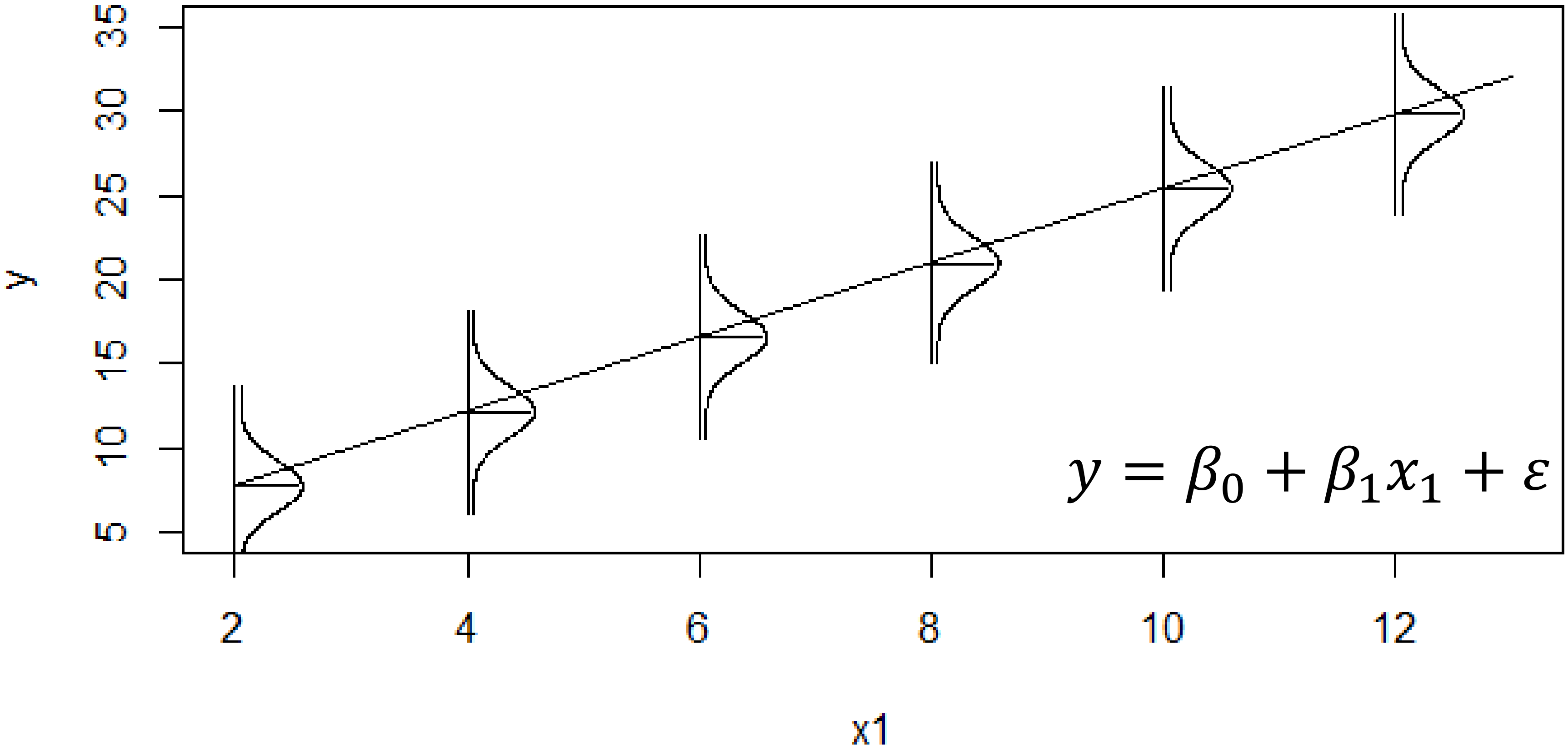
  β_0 is the value of y when each x_j equals zero

  β_j is the change in y based on a unit change in x_j

  ε ~ N(0, σ^2) and the ε’s are independent of each other
- Unless the situation being modeled is purely deterministic, there will be some random variability in the outcome. This random error, denoted by ε, is assumed to be normally distributed with a mean of zero and a constant variance (σ^2).
- Linear regression is doing well if all the variables are **uncorrelated.**

Example: Linear Regression with One Input Variable

- x_1 - the number of employees reporting to a manager
- y - the hours per week spent in meetings by the manager



$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Additional variables could be added to this linear regression model. For example, a categorical attributes, such as engineering, finance, manufacturing, or sales.

Representing Categorical Attributes

$$y = \beta_0 + \beta_1 employees + \beta_2 finance + \beta_3 mfg + \beta_4 sales + \varepsilon$$

Possible Situation	Input Variables
Finance manager with 8 employees	(8,1,0,0)
Manufacturing manager with 8 employees	(8,0,1,0)
Sales manager with 8 employees	(8,0,0,1)
Engineering manager with 8 employees	(8,0,0,0)

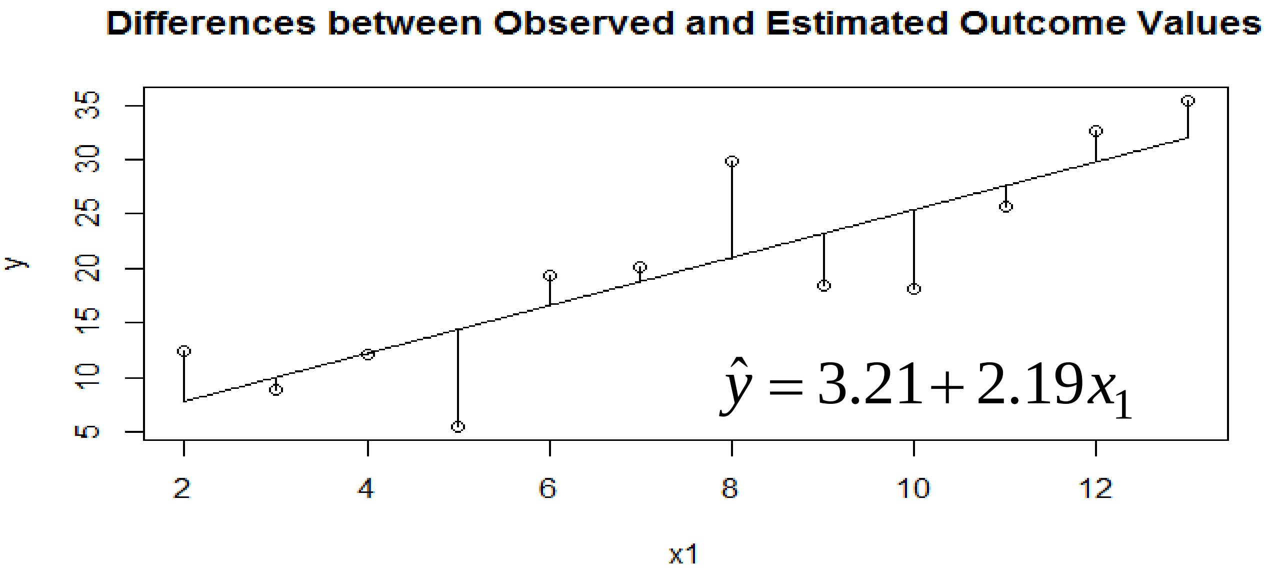
- For a categorical attribute with *m* possible values
  - Add *m-1* binary (0/1) variables to the regression model
  - The remaining category is represented by setting the *m-1* binary variables equal to zero

Representing Categorical Attributes

- Suppose it was decided to include the manager’s U.S. state of employment in the regression model.
  - Then 49 binary variables would have to be added to the regression model to account for 50 states.
  - Alternatively, it may make more sense to group the states into geographic regions or into other groupings such as type of location

Fitting a Line with Ordinary Least Squares (OLS)

- Choose the line that minimizes: $\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})]^2$
- A very common approach to determine the best fitting line is to choose the line that minimizes the sum of the squares of the differences between the observed outcomes in the dataset and the estimated outcomes based on the equation of the fitted line. This method is known as Ordinary Least Squares (OLS).
- Provides the coefficient estimates, denoted b_j



Interpreting the Estimated Coefficients, b_j

$$\hat{y} = 4.0 + 2.2 employees + 0.5 \, finance - 1.9 mfg + 0.6 sales$$

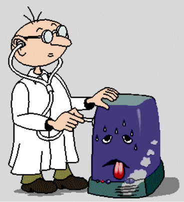
- Coefficients for numeric input variables
  - Change in outcome due to a unit change in input variable*
  - Example: $b_1 = 2.2$
    - Extra 2.2 hrs/wk in meetings for each additional employee managed*
- Coefficients for binary input variables
  - Represent the **additive difference from the reference level** *
  - Example: $b_2 = 0.5$
    - Finance managers meet 0.5 hr/wk more than engineering managers do*

* when all other input values remain the same

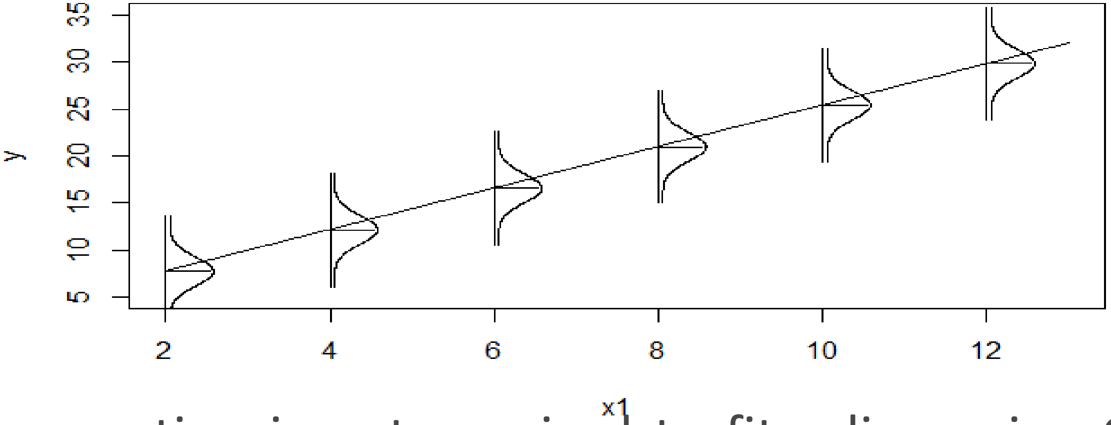
Interpreting the Estimated Coefficients, b_j

- Statistical significance of each coefficient
  - Are the coefficients significantly different from zero?
  - The used software package (R) performs a hypothesis test where the null hypothesis is that the coefficient equals zero and the alternate hypothesis is that the coefficient does not equal zero.
  - For small p-values (say <0.05), then the null hypothesis would be rejected and the corresponding variable should remain in the linear regression model. Small *p*-values means that the coefficient is statistically significant.
  - If a larger p-value is observed, then the null hypothesis would not be rejected and the corresponding variable should be considered for removal from the model.

Diagnostics – Examining Residuals

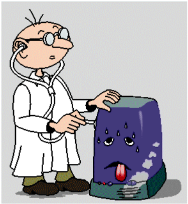


- Residuals
  - Differences between the observed and estimated outcomes
  - The observed values of the error term, ε, in the regression model
  - Expressed as: $e_i = y_i - \widehat{y}_i \qquad for\ i = 1,2\ldots,n$
- Errors are assumed to be normally distributed with
  - A mean of zero
  - Constant variance

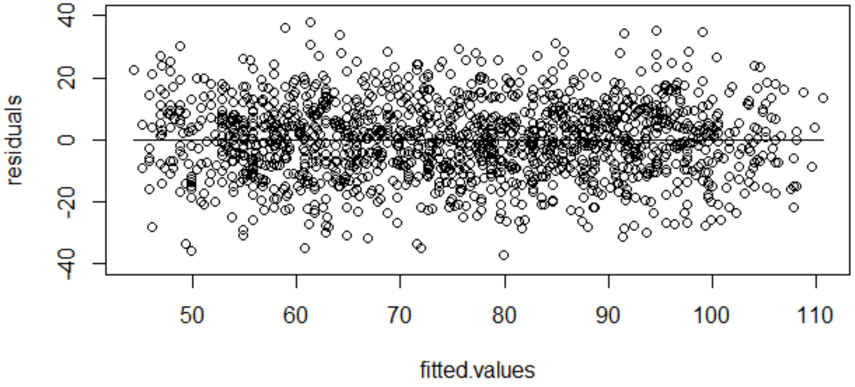


- Although this normality assumption is not required to fit a line using OLS, this assumption is the basis for many of the hypothesis tests and confidence interval calculations performed by statistical software packages

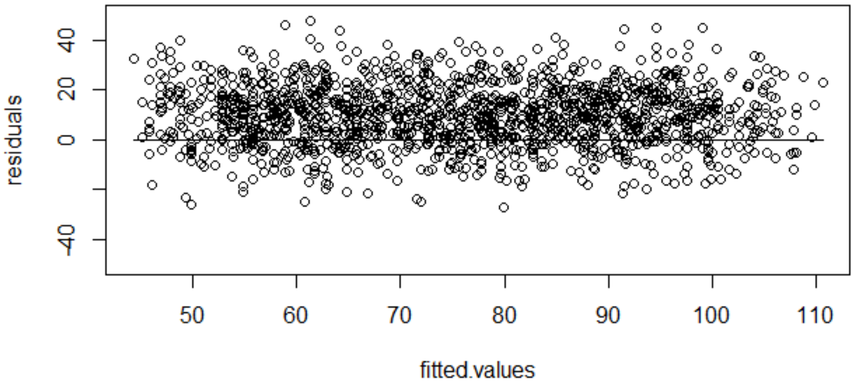
Diagnostics – Plotting Residuals



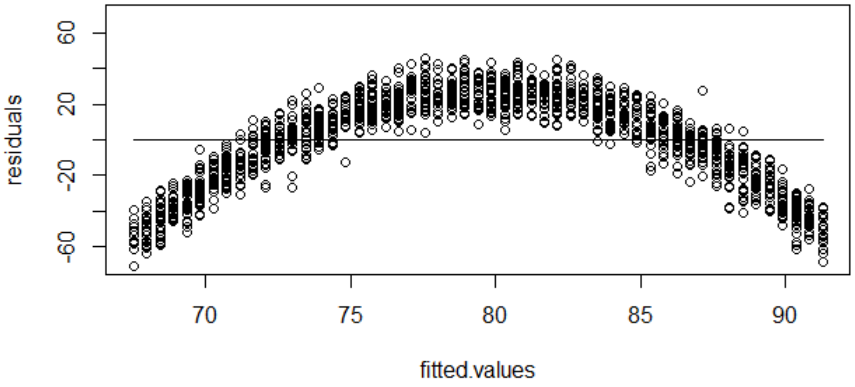
Ideal Residual Plot



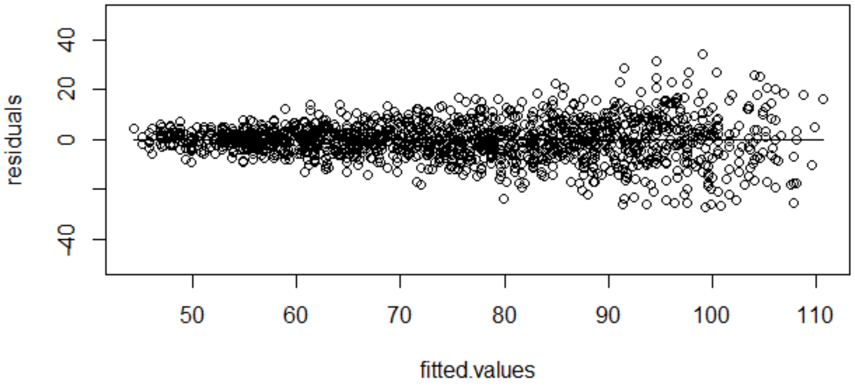
Non-centered



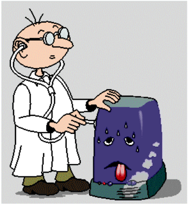
Quadratic Trend



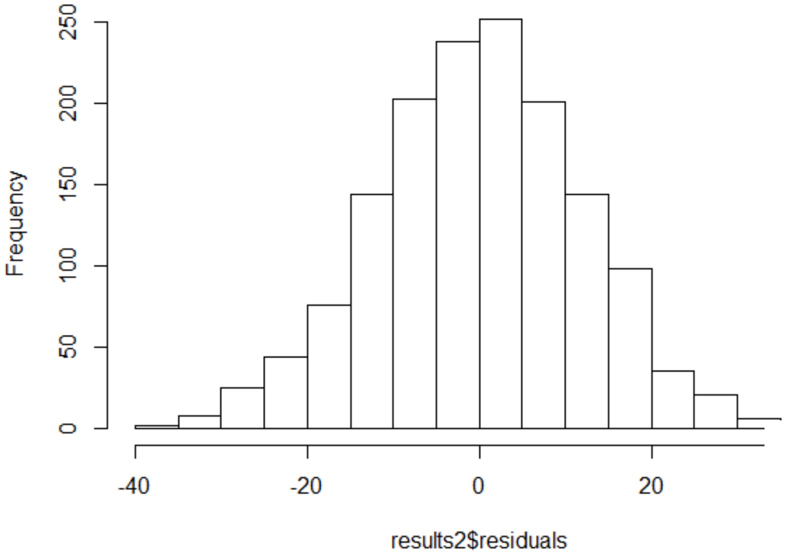
Non-constant Variance



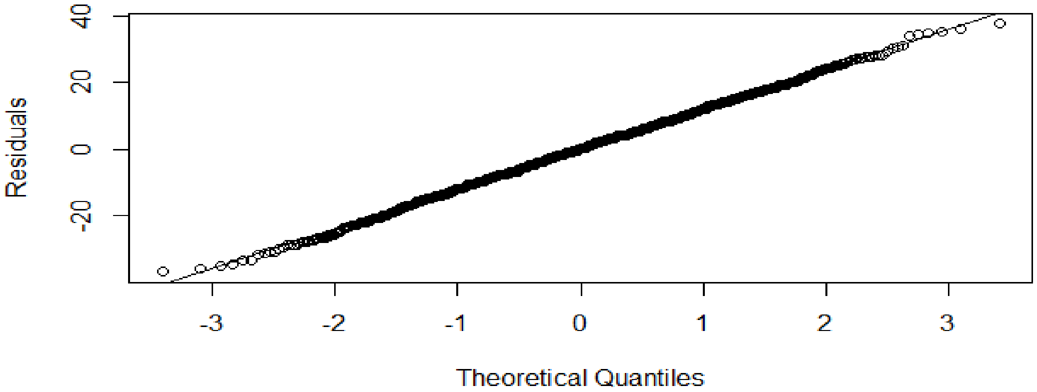
Diagnostics – Residual Normality Assumption



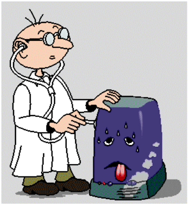
Ideal Histogram



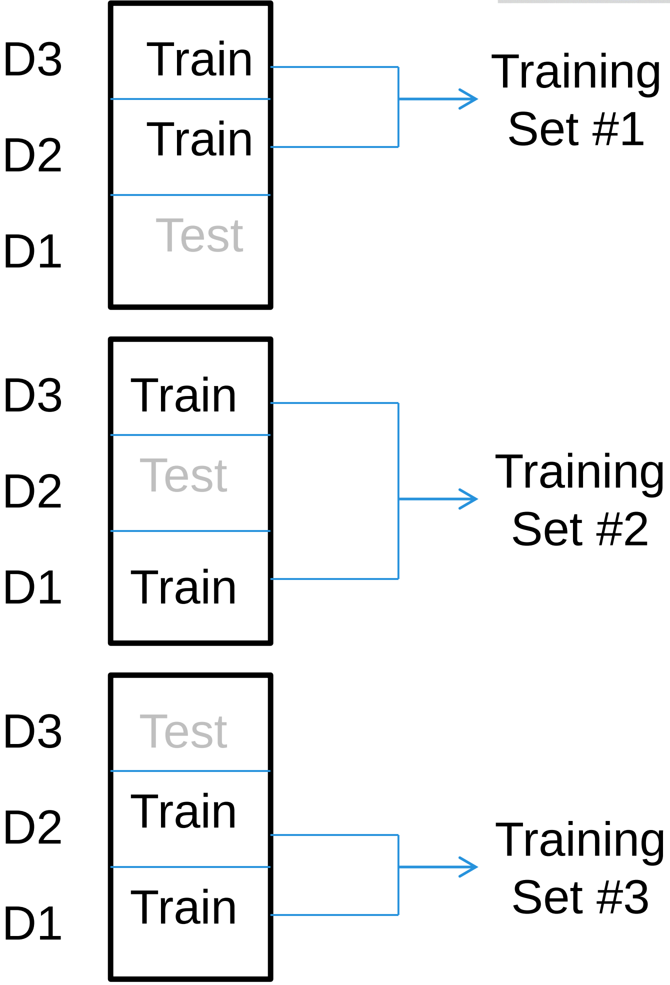
Ideal Q-Q Plot



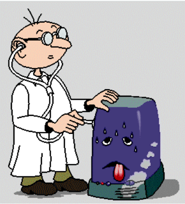
Diagnostics – Using Hold-out Data



- Hold-out dataset
  - Training and testing datasets
  - Does the model predict well on data it hasn't seen?
- N-fold cross validation
  - Randomly partition the data into N groups.
  - This method is used when you don't have enough data to create a hold-out dataset.
  - Holding out each group,
    - Fit the model
    - Calculate the residuals on the group
  - Estimated prediction error is the average over all the residuals.



Diagnostics – Other Considerations

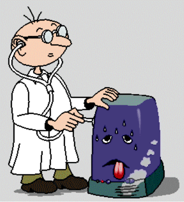


- R^2 metric
  - R^2 value quantifies goodness of fit: The fraction of the variability in the outcome variable explained by the fitted regression model.
  - $R^2 = 1 - SSerr/Sstot$

where SSerr = Sum[$(y-y_{pred})^2$] is the sum of the square of the errors and

SStot = Sum[$(y-y_{mean})^2$] is the sum of the square of the distances of the points from a horizontal line through the mean of all Y values.

Diagnostics – Other Considerations

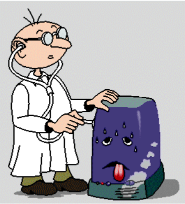


$R^2 = 1 – SSerr/Sstot$

where $SSerr = Sum[(y-y_{pred})^2]$ and $SStot = Sum[(y-y_{mean})^2]$

- For a good fit, we want an R2 value near 1.
- It is a fraction between 0.0 and 1.0, and has no units. Higher values indicate that the model fits the data better.
- When R^2 equals 0.0, the best-fit curve fits the data no better than a horizontal line going through the mean of all Y values. In this case, knowing X does not help you predict Y.
- When R^2=1.0, all points lie exactly on the curve with no scatter. If you know X you can calculate Y exactly.

Diagnostics – Other Considerations



- Identify correlated input variables, as regression modeling works best if the input variables are independent of each other
  - Pair-wise scatterplots
  - Check the coefficients
  - **If two input variables, *x_1* and *x_2*,** are linearly related to the outcome variable *y*, but **are also correlated to each other, it may be only necessary to include one of these variables in the model.**
    - Are the magnitudes excessively large?
    - Do the signs make sense?
  - Coefficients with large magnitudes or intuitively incorrect signs are also indications on correlated input variables.
  - Infinite magnitude coefficients could indicate a variable that strongly predicts a subset of the output

Linear Regression - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
The estimated coefficients provide a **concise representation** of the outcome variable as a function of the input variables.	Does not handle missing values well
Robust to handle redundant or correlated variables although it loses some explanatory value in the case of correlated variables	Assumes that each variable affects the outcome linearly and additively. The model will often not explain the data well. Variable transformations and modeling variable interactions can address this issue to some extent.
Explanatory value The estimated coefficients provide the explanatory value of the model and are used to easily determine how the individual input variables affect the outcome (Relative impact of each variable on the outcome)	Does not easily handle variables that affect the outcome in a discontinuous way Step functions
Easy to score data using coefficients	Does not work well with categorical attributes with a lot of distinct values For example, ZIP code

Check Your Knowledge

Your Thoughts?

1. How is the measure of significance used in determining the explanatory value of a driver (input variable) with linear regression models?
2. Detail the challenges with categorical values in linear regression model.
3. Describe N-Fold cross validation method used for diagnosing a fitted model.
4. List two use cases of linear regression models.
5. List and discuss two standard checks that you will perform on the coefficients derived from a linear regression model.

Module 4: Advanced Analytics – Theory and Methods

Part 3: Linear Regression - Summary

During this Part the following topics were covered:

- General description of regression models
- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The Reasons to Choose (+) and Cautions (-) of the linear regression model

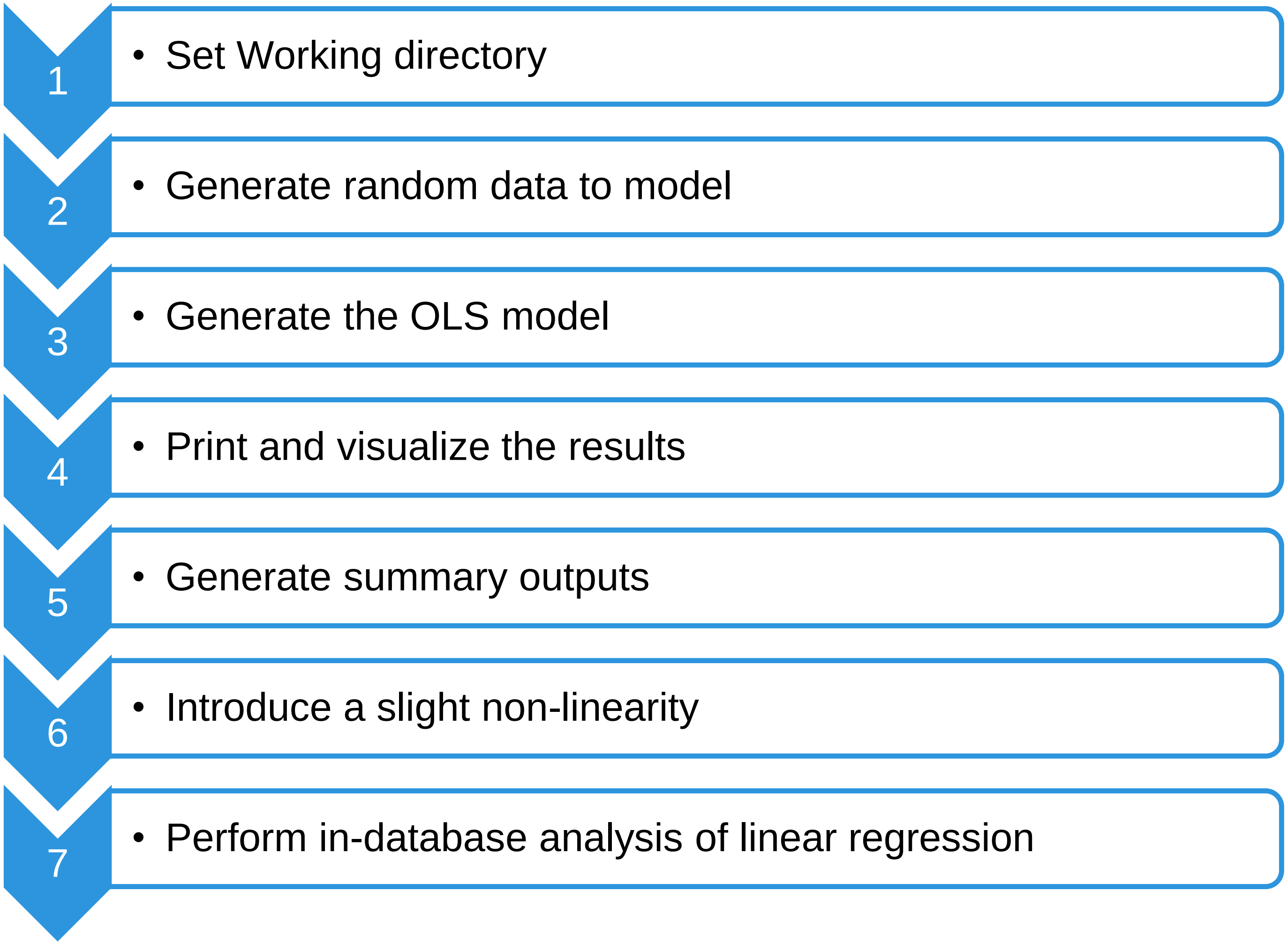
Lab Exercise 6: Linear Regression

This Lab is designed to investigate and practice Linear Regression.

After completing the tasks in this lab you should be able to:

- Use R functions for Linear Regression (Ordinary Least Squares – OLS)
- Predict the dependent variables based on the model
- Investigate different statistical parameter tests that measure the effectiveness of the model

Lab Exercise 6: Linear Regression - Workflow



1. Set Working directory
2. Generate random data to model
3. Generate the OLS model
4. Print and visualize the results
5. Generate summary outputs
6. Introduce a slight non-linearity
7. Perform in-database analysis of linear regression

Thanks